Project: Insights and Findings about Books

Group: Viz Wizards

I. Introduction

Reading is a beneficial habit that everyone should pick up on in their free time due to the vast amount of knowledge that a person can explore. However, beginners who want to take on this habit often do not know where to start because there are so many things to consider before actually buying a book. For instance, many beginner readers choose their first book based on how popular the book is, meanwhile others might make the decision based on the price. Some readers might just choose a book because they just wanted to form a habit of reading books and are willing to explore any genres out there. The key takeaway is that many beginners do not have a reference point to consider when choosing their first book or genre, and that is exactly what this project will address. Through comprehensive visualizations, we aim to deliver some reference points about books that are arguably more well-known so that not only beginners but readers of all levels can take them into consideration when choosing their books. Moreover, this project is also useful for editors who are interested in the topics of books and want to write articles about it (i.e: prices differences for site recommendations, popular books/genres over the years based on reviews/ratings and so on). It is also beneficial for authors who want to see the changes of books industry.

II. Data Analysis

The attributes in the dataset:

Attribute	Semantics	Attribute Type	Cardinality
title	Title of the book	categorical	222 unique values Has 366 data points
amazon_author	Author (first and last name) of the book	categorical	163 unique values Has 366 data points
amazon_rating	Rating of the book given by Amazon user on a scale of 1 to 5	quantitative	Range from 3.3 to 4.9 Has 366 data points

amazon_num_reviews	Number of written reviews of the book given on Amazon	quantitative	Range from 548 to 87841 Has 366 data points
amazon_price	Price of the book as of October 13, 2020	quantitative	Range from 0 to 105 Has 366 data points
amazon_year	Year the book was ranked on the bestsellers list	Ordinal	Range from 2009-01-01 to 2019-01-01 Has 366 data points
amazon_genre	Whether the book is fiction or non-fiction	categorical	2 levels (fiction or non-fiction) Has 366 data points
conlit_genre	Genre of the book, lists 1 genre out of 12 categories	categorical	10 categories Has 111 data points
conlit_pubdate	Original publication date of the book	Ordinal	Range from 2005-01-01 to 2017-01-01 Has 111 data points
conlit_author_gender	Gender of the author (M/F/O)	categorical	2 levels (F/M) Has 111 data points
conlit_author_nationality	Nationality of the author	categorical	9 unique values Has 65 data points
conlit_total_ratings	Total number of ratings of the book on Goodreads as of May 23, 2022	quantitative	Range from 39 to 4322160 Has 111 data points
goodreads_rating	Global Goodreads rating	quantitative	Range from 3.28 to 4.73 Has 366 data points
goodreads_series	Series name	categorical	Has 83 unique values Has 131 data points
goodreads_genres	Genre(s) of the book	categorical	212 unique values
goodreads_edition	Type of edition	categorical	Has 25 unique values Has 100 data points
goodreads_publisher	Publisher of book	categorical	Has 127 unique values Has 359 data points
goodreads_publish_date	Publication date	Ordinal	Range from 1970-01-01 to 2020-03-17 Has 359 data points

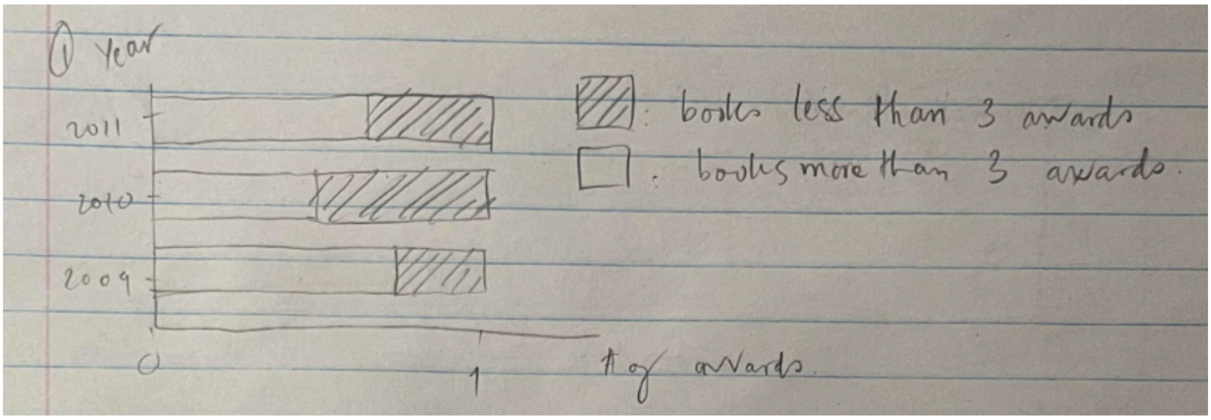
goodreads_first_publish_date	Publication date of first edition	Ordinal	Range from 1952-01-01 to … Has 176 data points
goodreads_awards	List of awards received by the book	categorical	611 unique values
goodreads_num_ratings	Number of total ratings on Goodreads	quantitative	Range from 608 to 6376780 Has 359 data points
goodreads_likedPercent	Percent of ratings over 2 stars on Goodreads	quantitative	Range from 74 to 98 Has 359 data points
goodreads_price	Price of the book (extracted from Iberibro)	quantitative	Range from 0.85 to 86.87 Has 275 data points
nyt_published_date	Date the list was published	Ordinal	Range from 2010-01-03 to 2019-12-29 Has 169 data points
nyt_list_name_encoded	Category of the list	categorical	Has 26 unique values Has 169 data points
nyt_price	Price of the book	quantitative	Range from 0 to 40 Has 169 data points
nyt_weeks_on_list	Number of weeks the book was on the best sellers list	quantitative	Range from 0 to 561 Has 169 data points

III. Task analysis

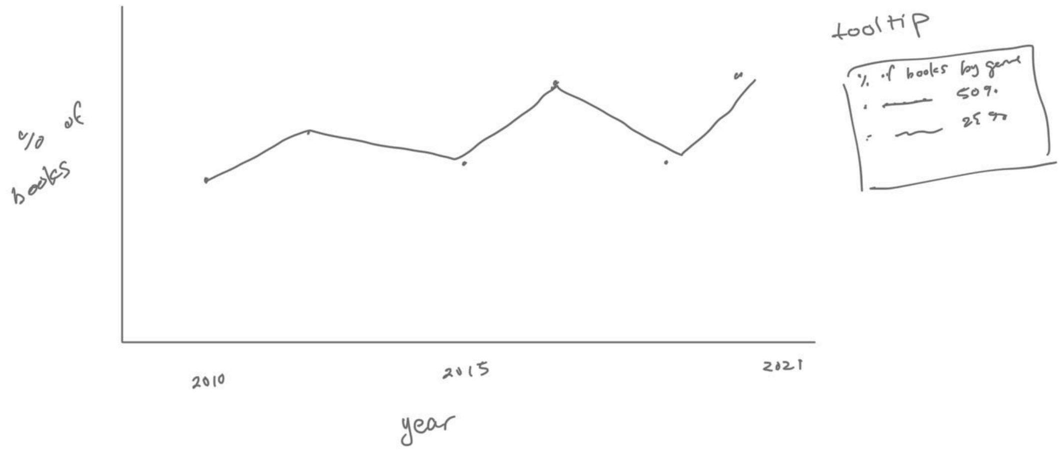
1. Which year has the highest proportions of books that have more than 3 awards? (Edison)
2. Most popular book genres over the years? (Tien)
3. What are the top book genres for top 5 publishers with highest occurrences? (Tien)
4. What years had the books with the best ratings? (Average ratings per year?) [Javier]
5. How do the prices differ between Goodreads and Amazon platforms per book listed in XYZ genre(s)? (Javier)
6. What is the number of books per rating filtered by numbers of reviews on Amazon and Goodreads? (gabrielle)
7. Do books with more awards tend to stay longer on the NYT's best selling list? (gabrielle)

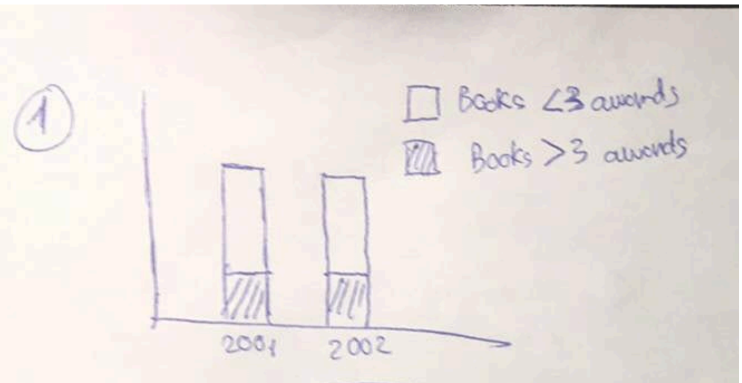
IV. Sketches

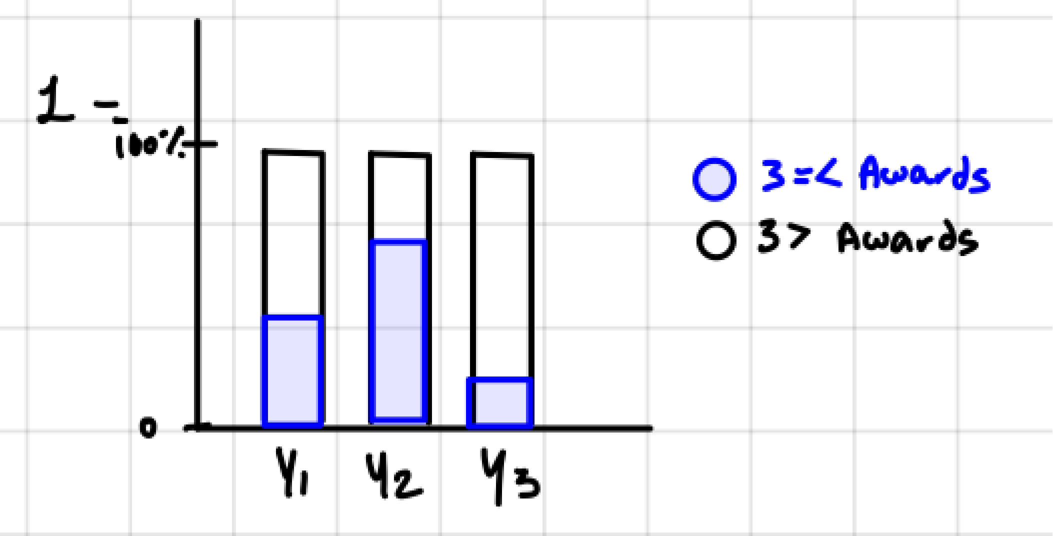
Task 1: line graph chosen



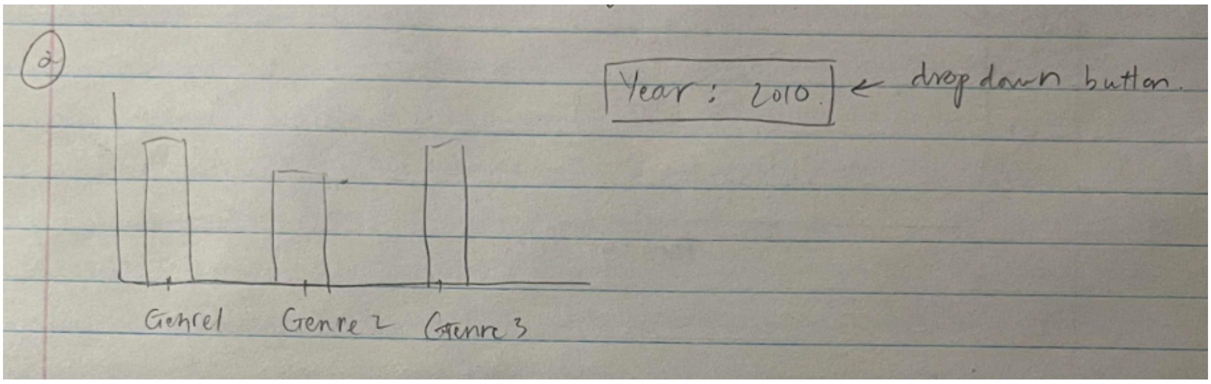
year of
Highest Proportion of books that have more than 3 awards

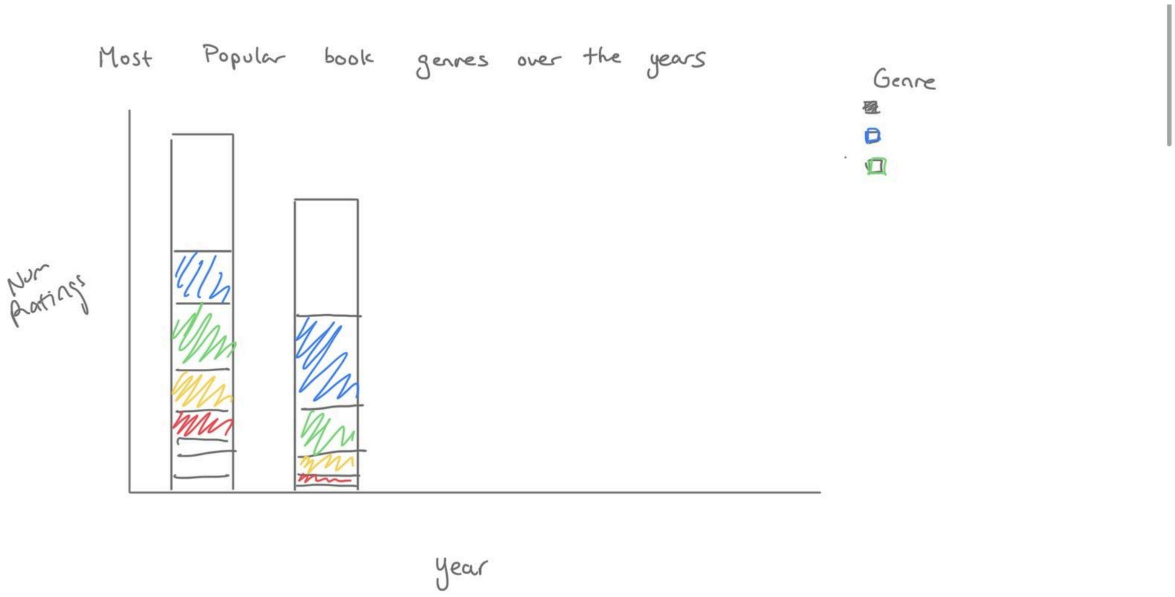


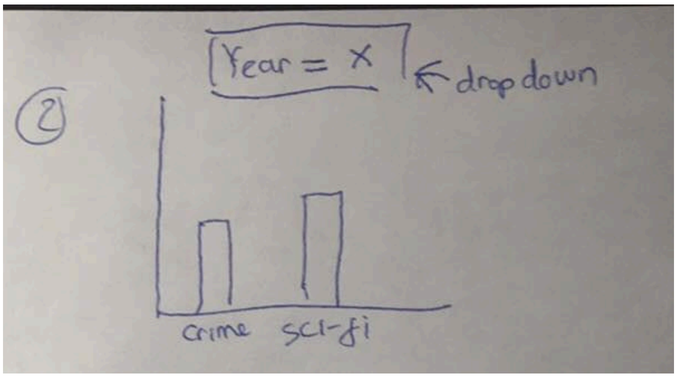


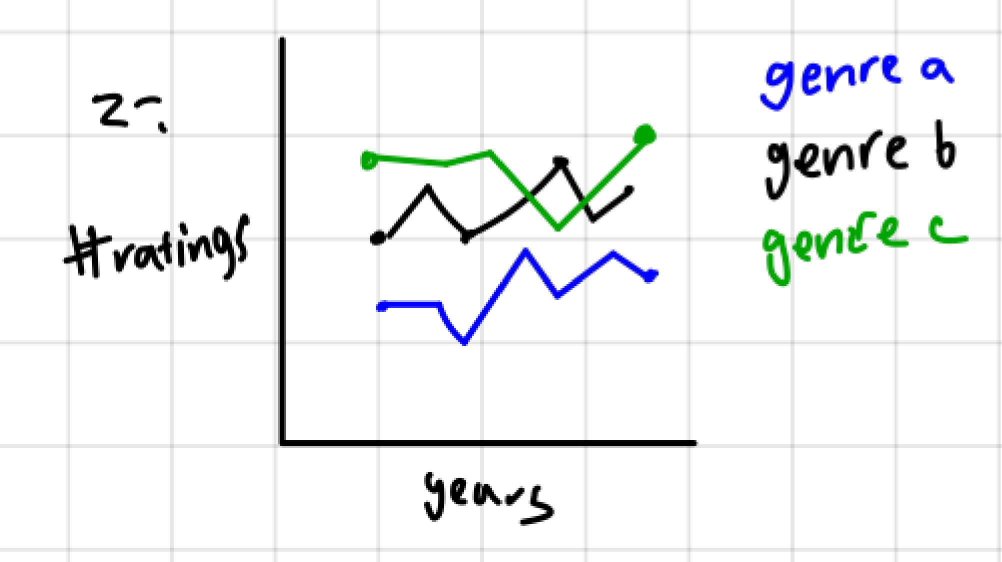


Task 2: Multi-view - Line and bar graph chosen

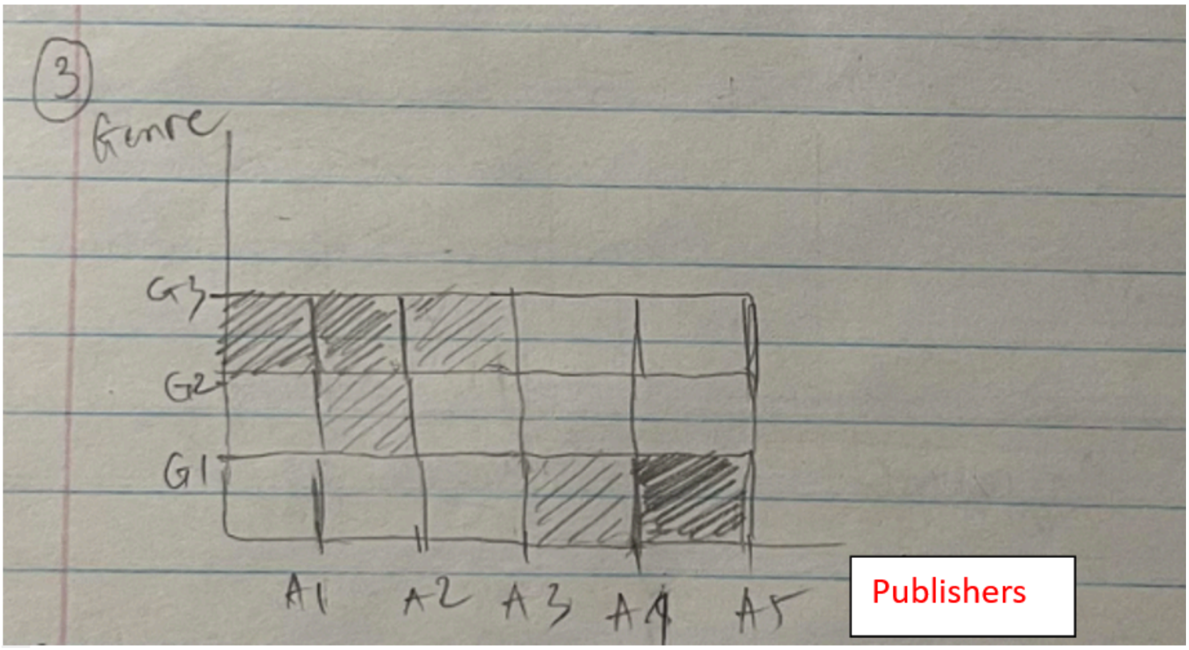


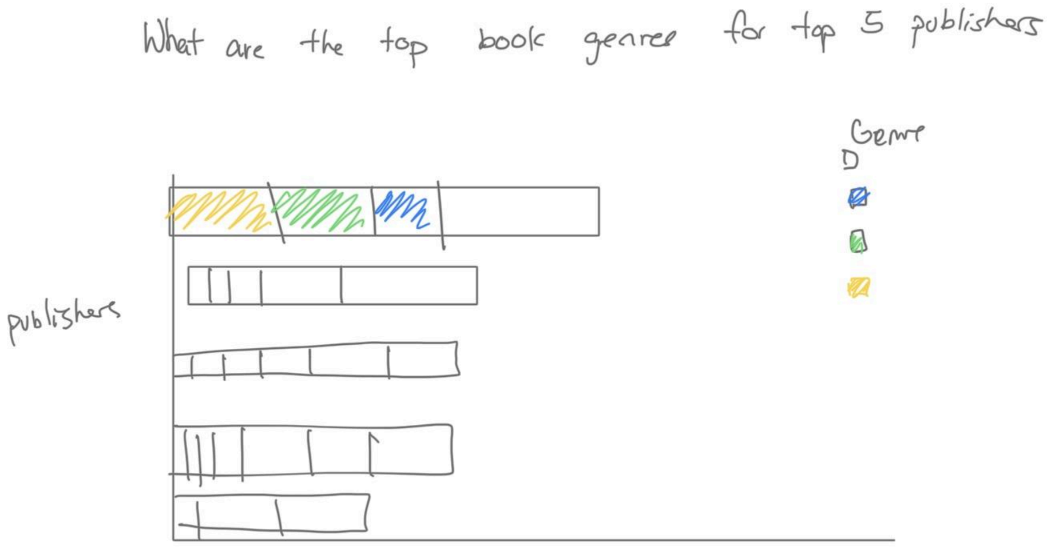


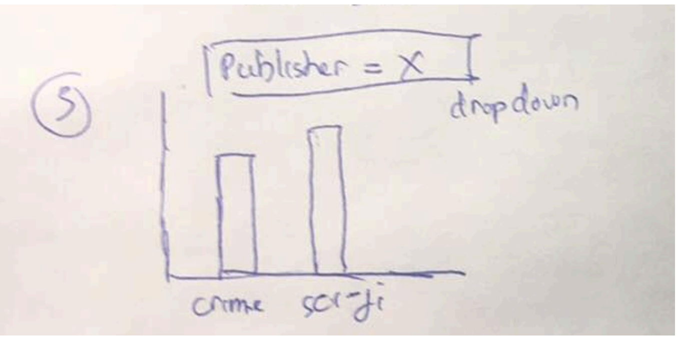


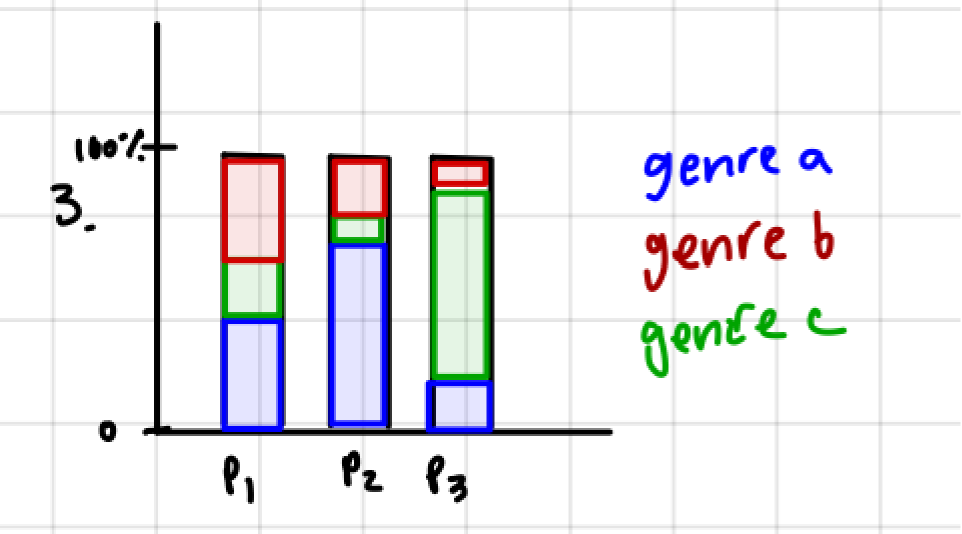


Task 3: Heatmap chosen

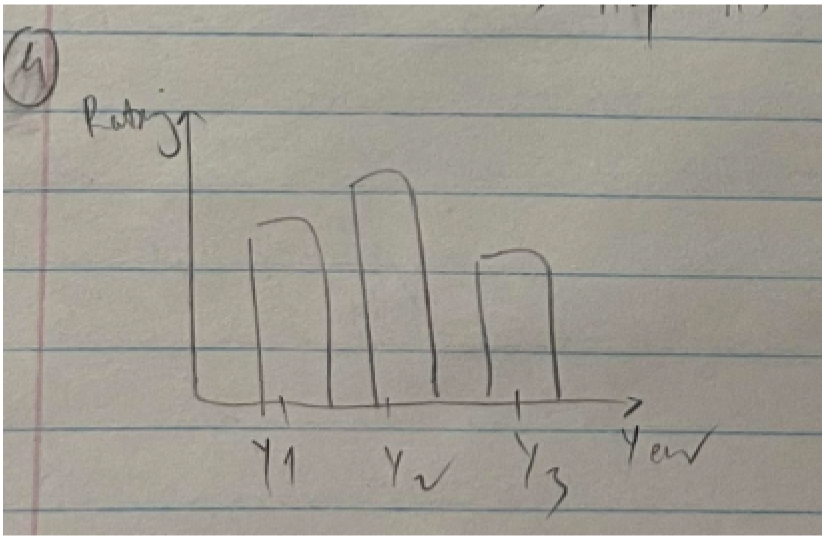




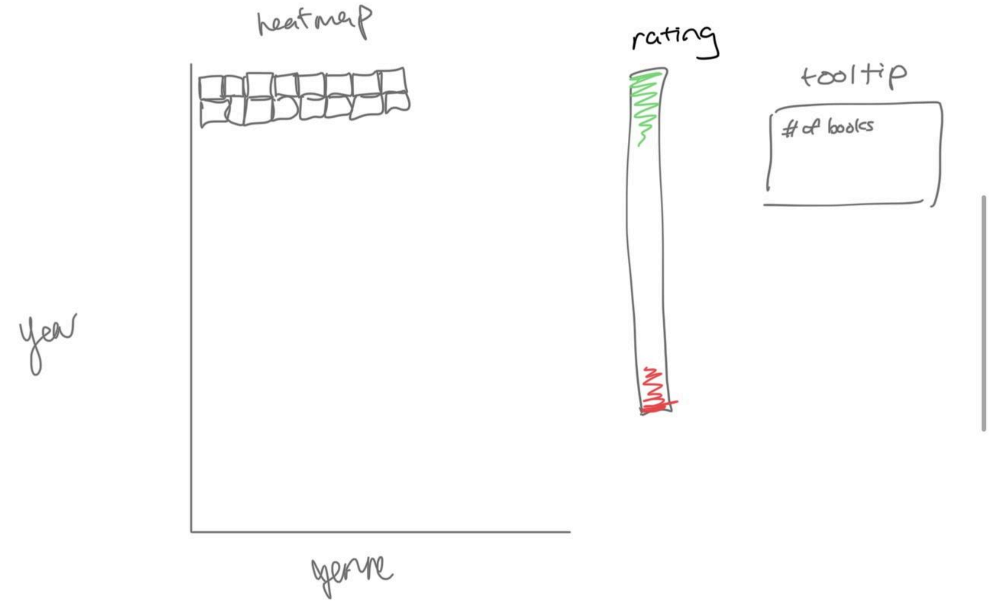


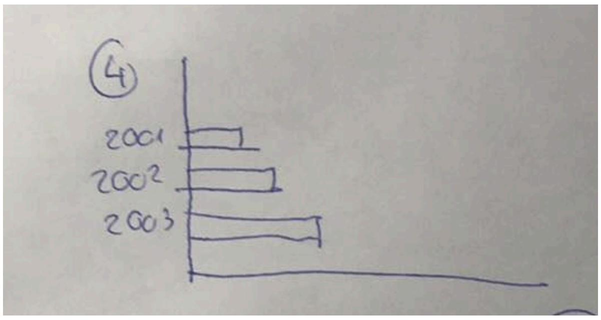


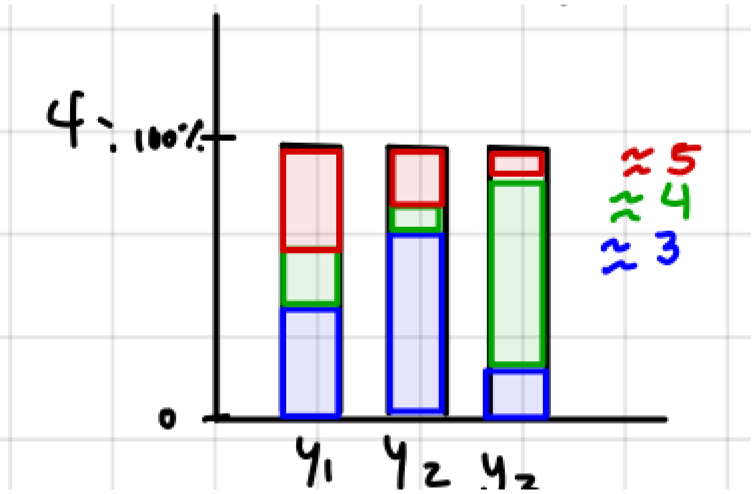
Task 4: Heatmap chosen



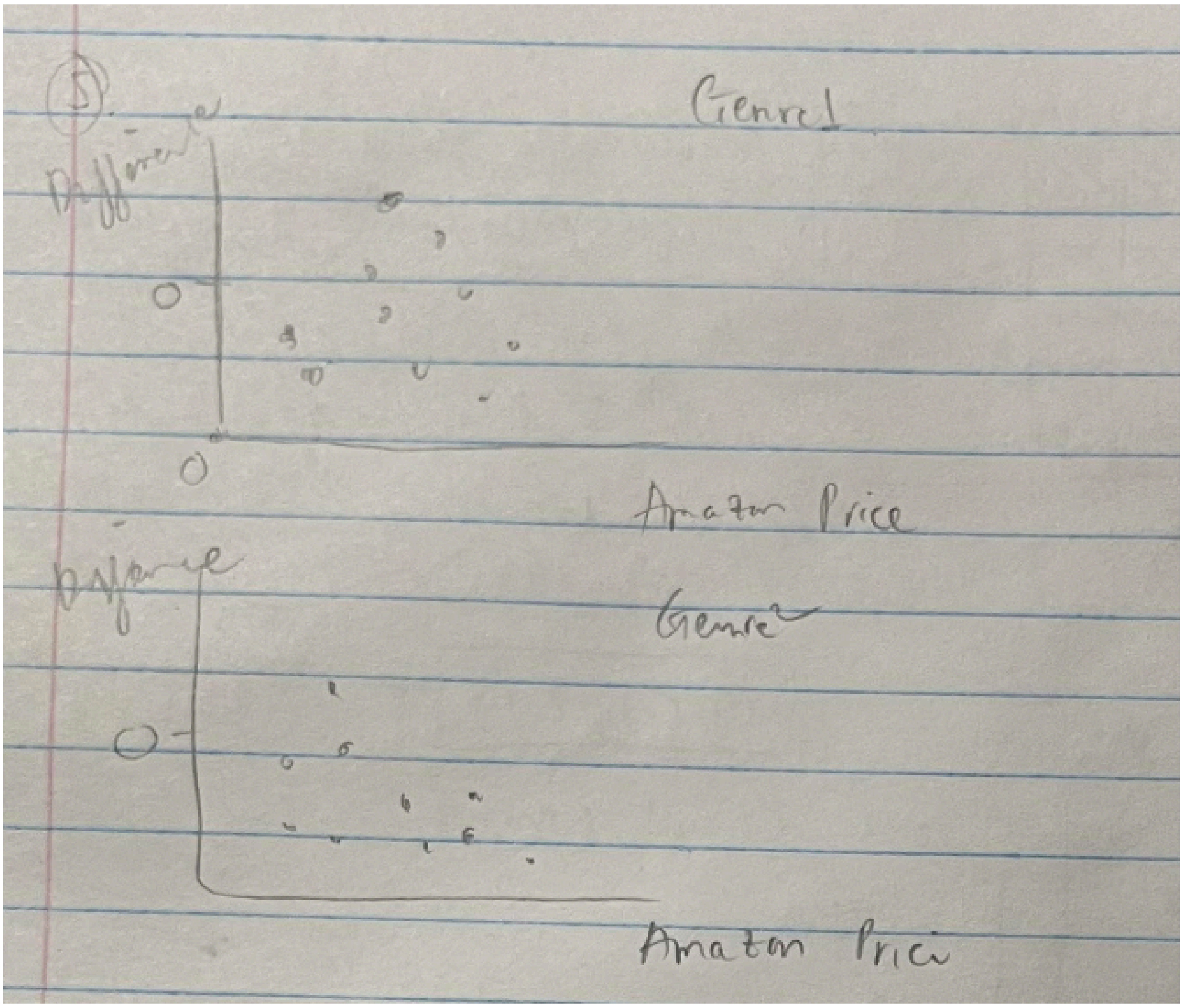
What years had the book with the best ratings

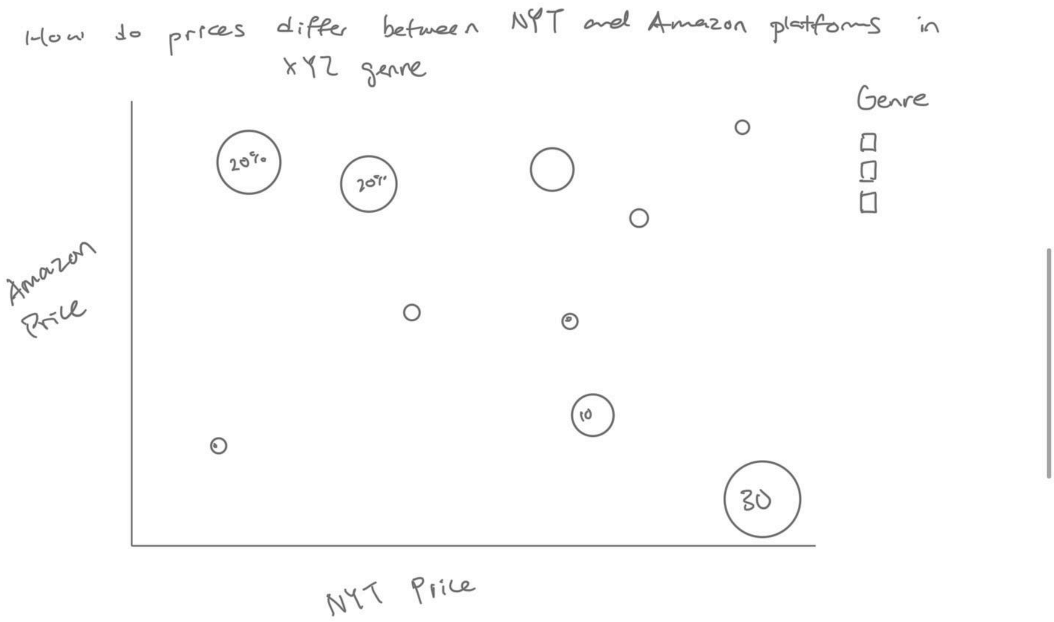


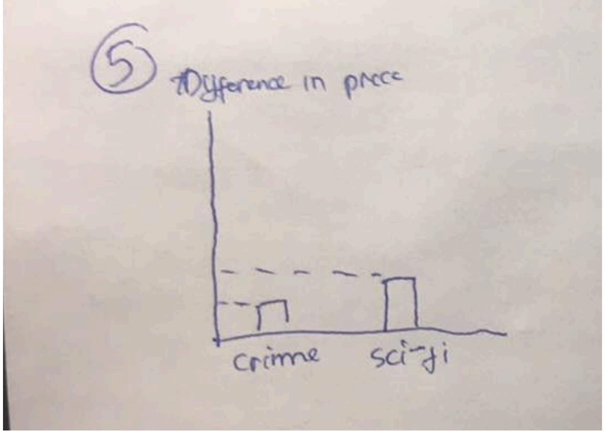


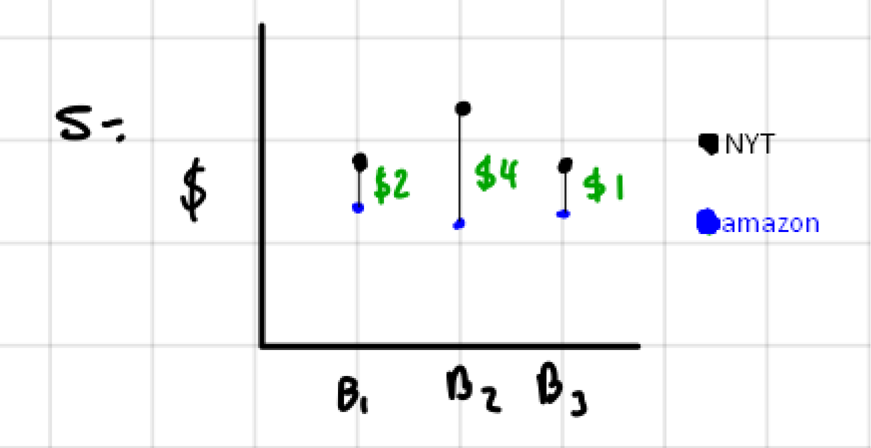


Task 5: Scatterplot with line connecting the two different prices chosen

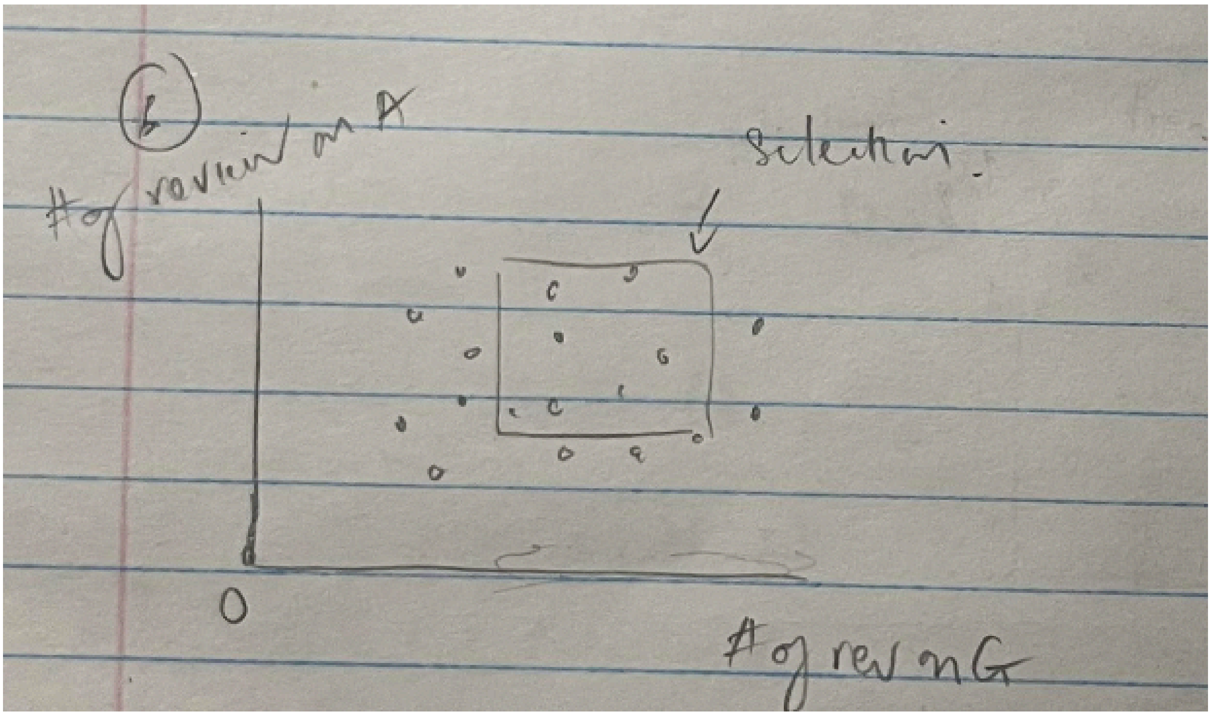




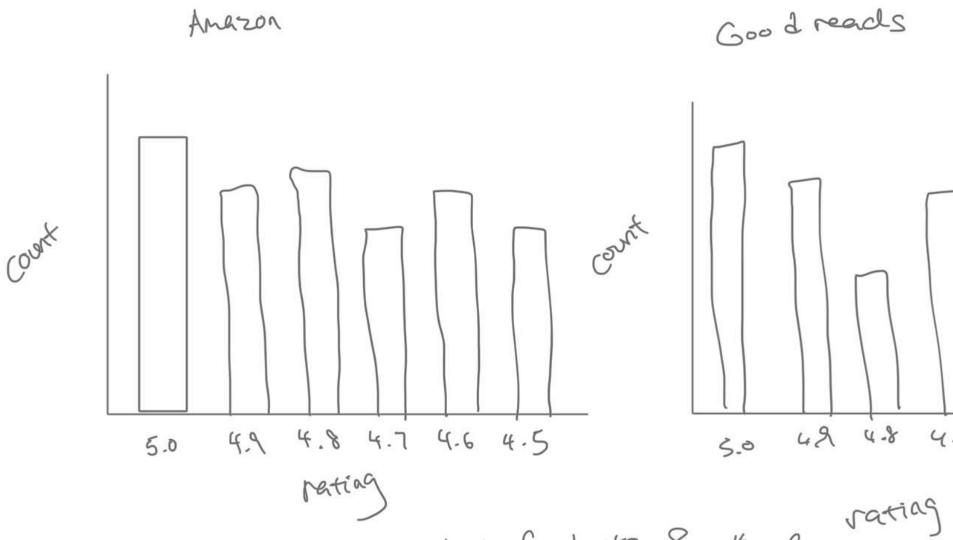




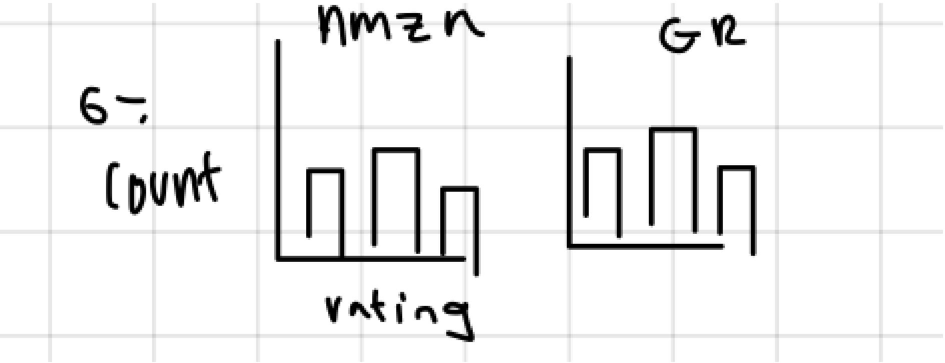
Task 6: Multi-view - scatter plot and bar graph chosen

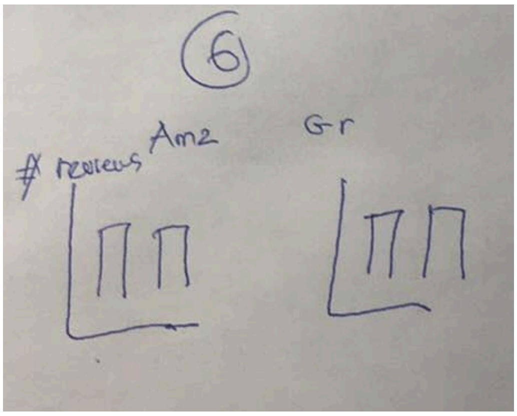


What are the top rated reviews

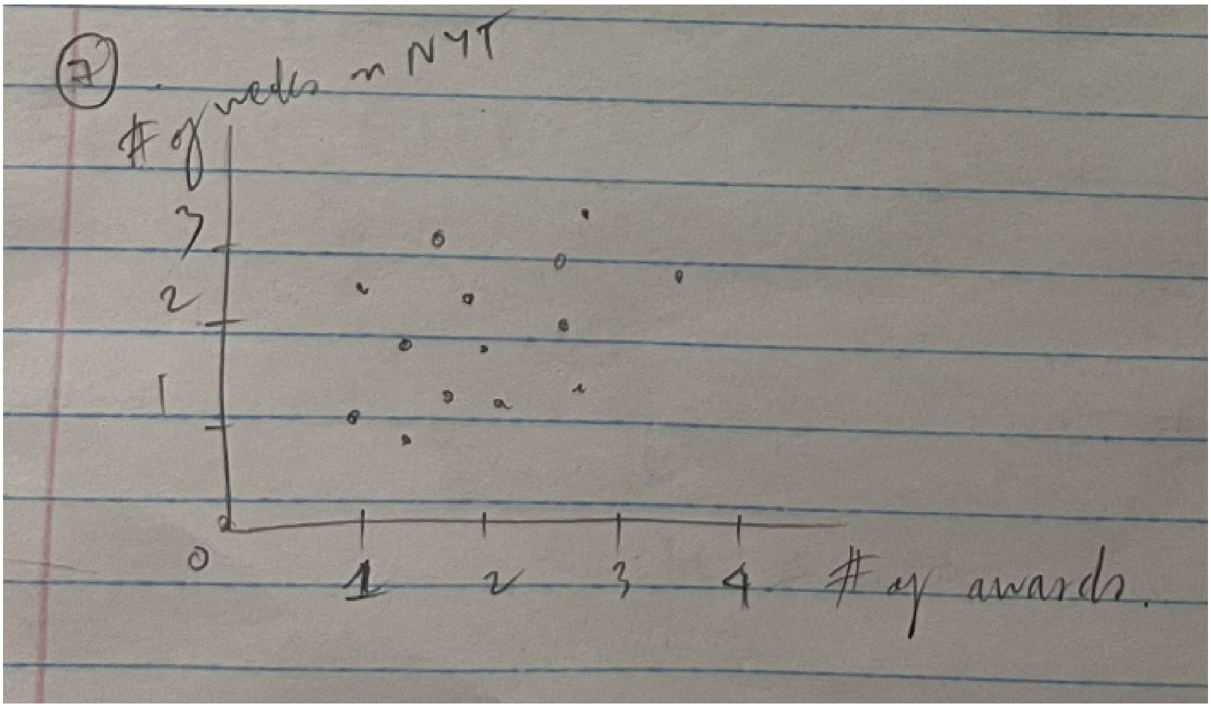


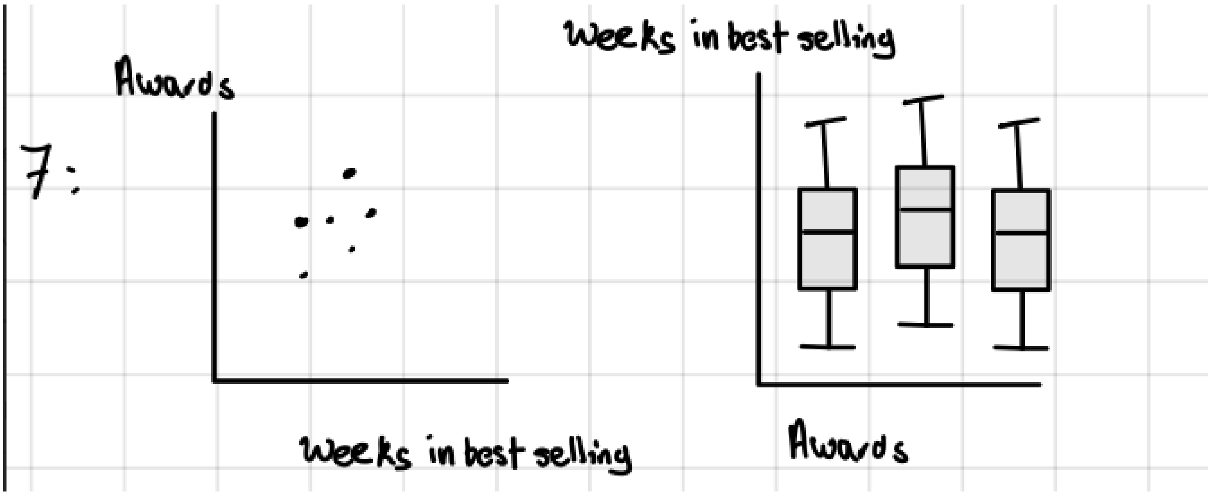
on click gives a list of books & # of ratings



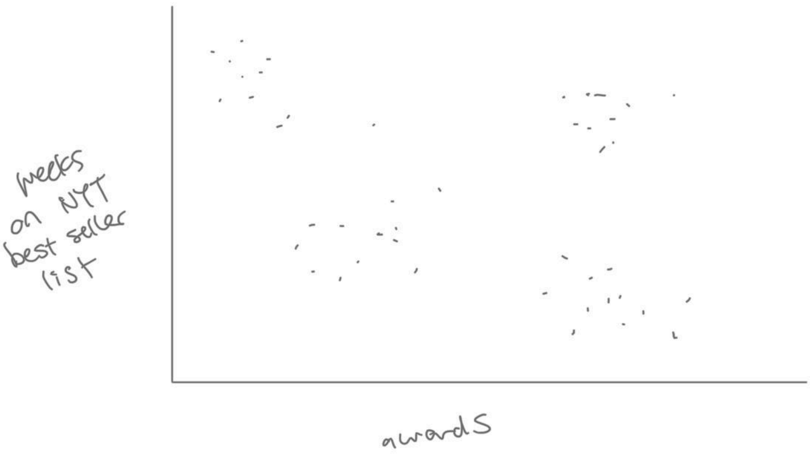


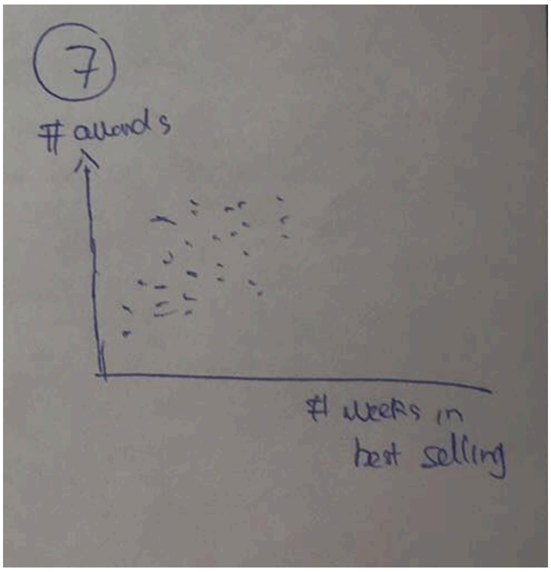
Task 7: Multi view - Scatter plot and box plot chosen





Do books with more awards tend to stay longer on NYT best selling list





V. Summary

Task 1: Line graph - Which year has the highest proportions of books that have more than 3 awards?
For this visualization, we wanted to show which year had the highest proportion of books that won more than 3 awards. The data we are using in this visualization is the proportion of books with more than 3 awards where the attribute type is quantitative and the publishing year of the book where the attribute type is ordinal. The channel used to encode the proportion of books is the vertical position and the mark used is points and the connection lines between them. We chose to use a line chart because the year is an ordinal attribute, so we thought it would be more appropriate to use a line chart rather than a bar chart. We also

decided we would highlight the highest point using a different colour so it can pop out and make it easier for us to locate. Another benefit is that we can observe whether there are any trends in books winning awards. However, since we used proportion we cannot access the count directly without any interaction and we can't tell how many of those books won exactly 4 awards.

Task 2: Multi-view (Line graph + Bar graph) - Most popular book genres over the years?

For this visualization we wanted to show the most popular genre of the books over the years. The data represented in the line graph is the proportion of books of each genre and the year that the book was published. For the line graph, the mark used is the points and the connection lines between them, for the channel we used vertical position to encode the percentage of books of a certain genre and horizontal position to encode the publishing year. The colour hue is the channel that encodes the type of genre since the type is categorical. For the bar graph, we used horizontal position to encode the count of the book and the vertical position to encode the genre. We will also sort the bar graph from highest to lowest starting from the top. We used a line graph because we thought we could show connectedness by connecting the dots together, and we also wanted to convey any trends that can be happening over time. For the bar graph, we decided on a common scale as the channel to encode the count of books of a certain genre because it would provide good accuracy and we can compare between the genres. We also sorted it so we can easily distinguish highest and the lowest counts, as well as compare between genres. We decided to use multiple views for this visualization because we can observe the overall trend in years and we can narrow the view to specific years as well.

Task 3: Heatmap - What are the top book genres for the top 5 publishers with the highest occurrences?

For the heatmap, the X and Y axes encode the publishers and the genres respectively. The color saturation will represent the number of books for each publisher-genre pair. Heatmap is useful here because we can visualize a common attribute from a combination of two categories (count of books for the combination of genre-publishers). For color, we are using a color palette (Viridis, or a similar palette) that allows for better readability (color deficiency audience included), and the palette also works well for a continuous attribute that has a magnitude or order. It will also have a legend that displays the color and the counts, as well as a mouse-hover tooltip to allow for the redundant encoding of the count so that the user can use the visualization interactively.

Task 4: Heatmap - What years had the books with the best ratings?

Similar to Task 3 above , we are using a heatmap for this task. X and Y axes encode the year and the genres respectively. The color will represent the number of books for each publisher-genre pair. The use of heatmap is useful because it allows us to visualize a common attribute from a combination of two categories (count of books for the combination of genre-year). The visualization will also have a legend that accompanies it to display the color and the counts, as well as a mouse-hover tooltip to allow for the redundant encoding of the count so that the user can use the visualization interactively. For color choices, we are using a color palette (Viridis, or a similar palette) that allows for better readability (color-deficiency included), and the palette also works well for a continuous attribute that has a magnitude/order which is appropriate as we are using the count of the genres.

Task 5: Points - How do the prices differ between Goodreads and Amazon platforms per book listed in XYZ genre(s)?
The idea is encoding the books' titles on the x-axis, and the price of these books in the y-axis, the points would represent the book in said platform, with a line connecting both points and a label that says the difference in dollars or percentage for each. The reason why we chose this plot is because it has great accuracy due to the common axis and we can use that common axis to compare price differences between books. We thought that adding a line between the two points would convey the price difference more effectively. Since there are far too many possible selections of books, we will add dynamic queries to filter books based on genres, publication years, authors and so on to limit the number of data points presented so that the plot would not appear too busy and overwhelming for users. Moreover, with the use of dynamic queries, users can filter and look at the books' price differences of their interest. However, this visualization cannot provide information about the average price differences in each genre.

Note: Originally, we wanted to compare Amazon prices vs NYT prices but after checking distribution of NYT prices, we see that most of the data points are 0 so we will use Goodreads prices instead. That's why in our sketches we have NYT prices vs Amazon prices.

Task 6: Multi-view (Scatter chart + Bar charts) - What is the number of books per rating filtered by numbers of reviews on Amazon and Goodreads?
In order to filter books based on the number of reviews on Amazon and on Goodreads, we decided to use a scatter plot with marks as circles representing each book, the number of reviews on Goodreads encoded on x-axis and the number of reviews encoded on Amazon on y-axis. This chart will be horizontally concatenated with two bar charts of count of books per rating on Goodreads and on Amazon, with ratings on x-axis and count of books on y-axis. Then we will use cross-filtering where we use an interval selection for x-axis and y-axis in the scatter plot to filter for the books with respect to the number of reviews on Amazon and Goodreads in the bar charts. With this method, we can allow freedom for users to choose ranges of the number of reviews they want to filter. Moreover, we used bar charts for count of books vs rating as we are comparing number of books for each rating and bar chart which has great accuracy due to common axis and makes it easy to find which rating has the highest count by comparing heights of the bars. Moreover, scatter plot and bar plot are easy to understand for a wide range of audience.

Task 7: Box Plot - Do books with more awards tend to stay longer on the NYT's best selling list?
Box plot is an efficient tool for representing the distribution in a numerical dataset, these are useful to compare distribution of the data across different categories or groups and to identify trends and outliers. In the case for this task, each box can represent the range of

awards, and the height or width of the box can represent the number of weeks on the list. In other words, the number of weeks on the best-selling list is encoded on y-axis and the number of awards is encoded on x-axis. Furthermore, box plot helps users to compare the distribution of number of weeks in best-selling per number of awards with high accuracy due to common axis. Correlation between number of awards and number of weeks on the best-selling list can be explored using boxplot. Lastly, users can easily spot which number of awards' distribution is straying from the trend or has some irregularities or outliers. However, if we wanted to know the shape of the distribution this visualization is not the most effective for it.